Journal of Emerging Computer Technologies
Research Article
Received: *2022-04-08* | ***Reviewing:*** *2022-04-08* & *2022-05-20* | ***Accepted:*** *2022-05-21* | ***Online:*** *2022-05-22* | ***Issue Date:*** *2022-06-30*

Ensemble Regression-Based Gold Price (XAU/USD) Prediction

Zeynep Hilal Kilimci
Department of Information Systems Engineering
Kocaeli University
Kocaeli, 41001, Turkey
zeynep.kilimci@kocaeli.edu.tr
0000-0003-1497-305X

***Abstract*—The prediction of any commodities such as cryptocurrency, stocks, silver, gold is a challenging task for the investors, researchers, and analysts. In this work, we propose a model that forecasts the value of 1 ounce of gold in dollars by using regression ensemble-based approaches. To our knowledge, this is the very first study in terms of combining regression models for the prediction of XAU/USD index although there are plenty of methods employed in the literature to forecast the price of gold. The contributions of this study are fivefold. First, the dataset is gathered between July 2019 and July 2020 from global financial websites in the world, and cleaned for modeling. Then, feature space is extended with technical and statistical indicators in addition to opening, closing, highest, lowest prices of gold index. Next, different regression and ensemble-based regression models are carried out. These are linear regression, polynomial regression, decision tree regression, random forest regression, support vector regression, voting regressor, stacking regressor. Experiment results demonstrate that the usage of stacking regression combination model exhibits considerable results with 2.2036 of MAPE for forecasting the price of XAU/USD index.**

Keywords—Gold price prediction, XAU/USD index forecast, ensemble regression, stacking regressor

I. INTRODUCTION

Gold, a precious metal, has maintained its popularity among societies for thousands of dec as a barter, reserve unit, and jewelry. Considering the durability of the gold mine due to its structure, the convenience it ensures in terms of its workability and other benefits, it is significant for the business in production and for the financial markets as a commodity. For this reason, the price of gold is widely followed in the world financial markets. If the gold index is the price of an ounce of gold traded in US dollars (XAU/USD), that is, it refers to how many US dollars it takes to buy an ounce of gold. Gold, as a commodity, is considered one of the most important investment instruments not only for companies that are in close contact with the outside world, but also for any country. Countries and multinational companies use the exchange rate, which is one of the most important economic variables, as well as gold reserves as variables to ensure their connection with the outside world. This makes the gold index and the gold market one of the largest and most important financial markets in the world. For this reason, the gold index can be quickly affected in a positive or negative way by many developments that may occur in the markets, the economy and political policies. Taking into account external factors, it becomes almost impossible to control the future level of the gold index and its market. This makes gold index forecasting a more attractive and active research area for researchers, and investors. Within the scope of this study, it is proposed to construct a model that predicts the price of the gold index based on the regression ensemble-based approach.

Regression analysis is known as a collection of statistical procedures for predicting the relations between a dependent argument and one or more independent arguments. Regression analysis is especially employed for two conceptively different objectives. Firstly, regression analysis is evaluated to conclude causal relations between the independent and dependent arguments. Second, regression analysis is commonly utilized for forecasting, where its use has drastically coincided with the area of machine learning. The second usage of regression analysis is the main focus of this work. The most used kinds of regression analysis are linear regression, logistic regression ridge regression, etc. In this study, linear regressor, decision tree regressor, random forest regressor, and support vector regressor are evaluated as base regressors on the other hand voting regressor, and stacking regressor are assessed as ensemble regression models to predict the gold price.

In this study, it is proposed to forecast the price of gold index (XAU/USD). Movements in XAU/USD are analyzed between July 2019 and July 2020 by gathering data from global financial websites. In order to compose feature set is, opening, closing, the highest, and the lowest gold prices are included to the dataset. The same variables of the dollar index that have an effect on gold have been added to the dataset. In order to extend feature space, technical indicators are also included namely, simple moving average, relative strength index, and Bollinger band. Then, five different regression models are constructed namely, linear regression, polynomial regression, decision tree regression, random forest regression, support vector regression. Finally, voting regressor and stacking regressor models are employed by consolidating previous four regression models to get more robust prediction of gold price. There are plenty of methods employed by academic circumferences to evaluate and forecast the price of gold, such models are based on linear regression (MLR), support vector machine (SVM), artificial neural network (ANN), etc. To our knowledge, this is the very first attempt in terms of combining regression models for the prediction of XAU/USD index. Experiment results demonstrate that the combining of regression models is an effective method to acquire more robust results for forecasting the gold price instead of employing individual estimates.

The rest of the article is organized as follows: Section 2 presents a summary of studies that analyze predict direction on financial investment instruments. Section 3 contains the architecture of the proposed model and the methods used for the construction of the system. The results of the experiment and the conclusions are presented in Section 4 and Section 5.

II. RELATED WORK

This part provides a summary of the literature studies on regression analysis to estimate the price or direction of

different instruments such as digital currencies, stocks, mineral commodities such as gold, silver, bonds, funds and such products.

In a study [1], authors propose to forecast gold prices using multiple regression method (MLR). Various parameters which have an impact on the gold prices are employed to construct feature set such as Commodity Research Bureau future index (CRB), USD/Euro Foreign Exchange Rate (EUROUSD), Inflation rate (INF), Money Supply (M1), New York Stock Exchange (NYSE), Standard and Poor 500 (SPX), Treasury Bill (T-BILL), and US Dollar index (USDX). Authors report that the success of proposed model is competitive for forecasting the price of gold with 85.2% of sample variations in monthly described by the model. In another study [2], authors aim to predict price of gold employing multiple linear regression with principal component analysis (PCA). In order to eliminate the problem of presence of multicollinearity of the explanatory variables, PCA is employed in the experiments. The usage of PCA contributes to the performance of the proposed system by improving prediction accuracy from 0.572 to 0.625. In other study [3], a new consolidated approach (ICA- GRUNN) based on independent component analysis (ICA) and gate recurrent unit neural network (GRUNN) is presented on the estimation of gold price. Authors conclude the paper that ICA-GRUNN outperforms the traditional techniques such as autoregressive integrated moving average (ARIMA), radial basis function neural network (RBFNN), long short-term memory neural network (LSTM), GRUNN, and ICA-LSTM in terms of accuracy.

In a study [4], authors investigate the impact of Chicago Board Options Exchange (CBOE) gold and silver implied volatility on gold futures volatility in China using heterogeneous autoregressive (HAR) and Ridge regression methods. To demonstrate the effectiveness of the proposed models, data is gathered between March 18, 2011 and June 29, 2018. Authors report that HAR and Ridge regression-based models perform better predictive performance compared to the benchmark models. Pierdzioch et al. present a real-time approach based on quantile-regression in order to assess forecast out-of-sample gold returns with the help of macroeconomic and financial parameters [5]. With the usage of real-time quantile-regression technique, model instability, uncertainty, and the possibility that a forecaster has an asymmetric loss function is provided. Authors inform that the forecasts calculated employing the real-time quantile-regression model performs better than an autoregressive model. In a study [6], Suranart et al. focus on the various models namely, neural network, radial basis function network and support vector regression (SVR) to forecast the price of gold. The dataset is collected between June 2008 and April 2013 and evaluated as monthly and weekly. Experiment results indicate that SVR exhibits superior performance on both weekly and monthly data by performing 1.140 of MAPE and 0.908 of MAPE, respectively. In a study [7], authors investigate the effect of decision tree and support vector regression models on the prediction of gold price. The decision tree technique is utilized for the feature selection task while the regression is performed for the purpose of gold index estimation. Experiment results show that the consolidation of decision tree and support vector regression models boost the prediction performance compared to the techniques namely, linear regression and neural network. Sadorsky proposes tree-based classifiers to forecast the direction of gold and silver price [8]. For this purpose, bagging, stochastic gradient boosting, and random forests are employed. The prediction performance of tree-based methods that ranges from 85% and 90% excels when compared to the logit models for 20-day and 10-day estimates. In another study [9], Mithu et. al propose to estimate the gold price employing regression methods for stock market inconsistency and settling economic. For this purpose, authors evaluate support vector regression (SVR), random forest regressor (RFR), decision tree, gradient boosting, and XGBoost methods separately to estimate the daily gold price. The paper is concluded the superior performance of RFR algorithm with 99% of accuracy score. In [10], authors focus on estimating artificial neural networks of gold prices using multiple linear regression models. Experiment results demonstrate that MRL analysis in an effective way to compose the network pattern and get more accurate estimation score with 0.004264% of MSE.

III. METHODS

In this part, regression and ensemble regression-based models are introduced used in this work.

A. Linear Regression (LR)

In statistics, linear regression is a linear model for constructing the relation between a scalar answer (dependent) and one or more expository parameters (independent). The state of one expository parameter is named as simple linear regression while the situation is named as multiple linear regression for more than one variable [11]. However, this is rather different from multivariate linear regression, where multiple related dependent parameters are estimated, rather than a one scalar parameter [12]. The correlations are composed of employing linear estimator functions whose unknown model variables are predicted from the data in linear regression. These are called as linear models [13]. Linear regression is the first and broadly employed type of regression analysis in various applications [14]. These are trend estimation in time series data, epidemiology, finance, economics, environmental science, machine learning, etc. Because linear regression is extensively applied in social, behavioral, biological and sciences to define possible relations between parameters, it is as one of the most significant tools employed in these fields.

B. Polynomial Regression (PR)

Polynomial regression is a form of Linear regression employed as a specific version of multiple linear regression which forecasts the relation between independent and dependent variables as an nth degree polynomial [15]. Even though polynomial regression complies a nonlinear model to the data, as a statistical forecast problem it is linear, that is the regression method is linear in the unknown variables that are forecasted from the data. That is why polynomial regression is accepted to be a specific situation of multiple linear regression.

C. Decision Tree Regression (DTR)

This approach is also known as decision tree learning or induction of decision trees which is one of the predictive modelling techniques used in machine learning, data mining, and statistics. It utilizes a decision tree as a predictive model to move from samples about an item demonstrated in the branches to outcomes about the item's goal value indicated in the leaves. In tree models, if the goal parameter will take a discrete set of values are named as classification trees while

goal variable will get continuous values are entitled as regression trees [16]. Decision trees are assessed as intelligibly and simple among the other widely-used machine learning techniques [17].

D. Random Forest Regression (RFR)

Random decision forests or random forests are supervised learning techniques that are utilized for both classification and regression. The term forest stands for lots of decision trees. The model builds an ensemble of decision trees at training phase and blends their decisions by making inferences about the goal category (classification) that is the highly voted by the community of decision trees or average estimation (regression) of the base trees. Random forest approach ensures superior predictive accuracy, and do not enable overfitting if there are adequate trees in the forest [18], is proposed by Breiman [19]. In this work, random forest algorithm is employed with 100 estimators.

E. Support Vector Regression (SVR)

Support vector machine (SVM) is initially presented by Vapnik [20]. Support vector classification (SVC) and support vector regression (SVR) are two major categories for SVMs. SVM is a learning framework employing a high dimensional feature space. It provides functions of prediction that are enlarged on a sub set of support vectors. A variant of a SVM for regression model is asserted in 1997 by Vapnik et al. [21]. This model is named as support vector regression (SVR). The method generated by SVC just bounds up with a sub set of the training data, for the cost function for constructing the method does not attach importance to training data that reach out further the margin. Likewise, the model generated by SVR solely is contingent on a subset of the training data, inasmuch as the cost function for establishing the system disregards any training data that is near to the model estimation. SVR is the widely applied version of SVMs. One of the basic properties of SVR is that in place of reducing the error of observed training, SVR proposes to decrease the generalized error bound so as to accomplish generalized success. The generalization error is composed of the consolidation of a regularization term that checks the complexity of the hypothesis space and the error of training.

F. Voting Regressor (VR)

A voting ensemble or majority voting is an ensemble machine learning technique that consolidates the estimations from a lot other method. This generally is employed to develop the success of the system by integrating various models instead of using any single learning technique. In other words, majority voting is utilized for both classification or regression, is performed by consolidating the decisions from multiple methods. In the event of regression [22], it is calculated by taking average of the estimations of all models, while the predictions for each class are aggregated and the final decision on class with the majority vote is estimated in classification [23]. In summary, a voting regressor used in this work is an ensemble meta-predictor that conforms distinct individual regressors, each on the entire dataset. Then, the average of individual forecasts is ensured to decide a final estimation.

G. Stacking Regressor (SR)

Stacking is an ensemble technique that employs a meta-learning method to learn how to best consolidate the estimations from two or more individual machine learning techniques [24]. The advantage of stacking approach is that it can benefit the abilities of a set of well-performed methods on a regression or classification task and make estimations that have better success than any individual method in the ensemble. The meta-model is frequently basic, maintaining a smooth explication of the estimations performed by the individual models. For this reason, linear methods are constantly utilized as the meta-model. For instance, linear regression is employed for regression tasks while logistic regression used for classification tasks for the purpose of forecasting a label of class.

H. Proposed Framework

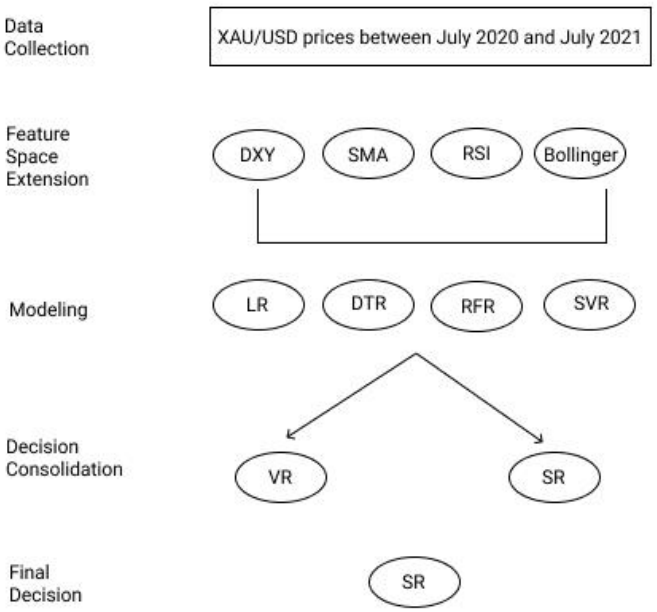

Fig. 1. The architecture of proposed model.

In this work, the prediction of gold price index (XAU/USD) is proposed. First, data collection process is performed between July 2019 and July 2020. Due to weekends and public holidays, 252 days of data were collected from the financial website. Missing 113 days of data is completed with a quadratic decal calculation method. Prices of opening, closing, highest, and lowest XAU/USD index are collected via the investing.com website with the help of Selenium library. The feature space of the data set is extended by collecting different features that may affect the direction and price of the gold index in the same date range. These are simple moving average (SMA) for 20, 50, and 100 days, opening, closing, highest and lowest dollar index (DXY) prices, 14-day relative strength index (RSI), the upper, middle and lower values of the Bollinger band (BB). SMA-20, SMA-50, SMA-100 mean 20-day, 50-day, 100-day simple moving averages of 1 ounce of gold, respectively. RSI-14 represents the 14-day relative strength index of 1 ounce of gold. BB-upper, BB-middle, BB-lower value symbolize the upper, middle, and lower band values of the Bollinger band of 1 ounce of gold. Historical indicator values are calculated using the technical analysis library named TA-lib. After extending feature space, modelling stage is implemented by utilizing various regression models namely, linear regressor (LR), decision tree regressor (DTR), random forest regressor (RFR), support vector regressor (SVR). To consolidate decisions of each regression model, voting (VR) and stacking regressors (SR) employed for the purpose of

acquiring more robust decision on predicting XAU/USD index. Figure 1 demonstrates major stages of tasks carried out along this work.

IV. EXPERIMENT RESULTS

In this section, we propose to forecast the value of 1 ounce of gold in dollars with the help of regression ensemble-based approaches. Experiment results of base regression models and ensemble regression methods are presented to demonstrate the impact of ensemble-based regression models on estimating XAU/USD index. Voting regressor (VR) and stacking Regressor (SR) are utilized as decision integration approach while linear regression (LR), polynomial regression (PR), decision tree regressor (DTR), random forest regressor (RFR), and support vector regressor (SVR) are evaluated as individual learners. The dataset is randomly splitted into 80% training and 20% test sets. To assess the performance of the proposed model, various evaluation metrics are employed namely, mean absolute percentage error (MAPE), mean absolute error (MAE), mean squared error (MSE), and R-squared (R^2) as below:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_a - y_f}{y_a}\right| \times 100 \quad (1)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_a - y_f| \quad (2)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_a - y_f)^2 \quad (3)$$

$$R^2 = 1 - \frac{\sum(y_a - y_f)^2}{\sum(y_a - y_m)^2} \quad (4)$$

where N is the size of test set, y_a and y_f are the actual and forecasted observations, respectively. In R-squared calculation, numerator corresponds to sum squared regression that is the sum of residuals squared while denominator corresponds to total sum of squares which the sum of the distance the data is away from the mean all squared. The statistical characteristics of the data are given followingly: The minimum value of data is 1404.00, the maximum value is 1810.78, the mean is 1574.508, and median of the data is 1547.00.

In Table I, experiment results of individual and ensemble-based regression techniques are presented in terms of different evaluation metrics. It is obviously observed that stacking regressor with 2.2036 of MAPE outperforms both the other ensemble-based regression model and base regression approaches. It is followed by SVR, PR, and LR with 2.2745, 2.3042, and 2.3816 of MAPE results. In this case, the usage of SVR, PR, or LR as an individual learner is more meaningful rather than the utilization of voting ensemble model in terms of both predict performance and time. On the other hand, voting regressor as an ensemble-based approach exhibits poor MAPE result compared to LR, PR, and SVR techniques. Moreover, it is clearly observed that the utilization of both DTR with 5.1923 of MAPE and RFR with 4.2544 of MAPE is not convenient to forecast the price of XAU/USD index. SVR as an individual method with 2.2745 of MAPE is competitive for forecasting of XAUSD index when compared to the proposed decision consolidation method, namely SR. Also, the utilization of decision tree-based models namely, decision tree and random forest regressors are not convenient to predict the price of XAUUSD when considering both system success and time.

TABLE I. EXPERIMENT RESULTS OF INDIVIDUAL AND ENSEMBLE-BASED REGRESSION MODELS

	Evaluation Metrics			
Models	**MAPE**	**MAE**	**MSE**	**R^2**
LR	2.3816	0.0198	0.0022	0.9975
PR	2.3042	0.0182	0.0021	0.9977
DTR	5.1923	0.1396	0.1275	0.9861
RFR	4.2544	0.0213	0.1065	0.9905
SVR	2.2745	0.0116	0.0018	0.9980
VR	3.1004	0.0174	0.0043	0.9969
SR	**2.2036**	0.0109	0.0011	**0.9986**
Avg.	**3.1016**	0.0341	0.0350	0.9950

The inclusion of stacking regressor model ensures roughly 3% improvement while voting regressor provides nearly 2% enhancement in terms of mean absolute percentage error compared to the poorest technique, namely DTR. As a result of Table I, the performance order of models can be summarized as: SR> SVR> PR> LR> VR> RFR> DTR. The success order of all models is similar compared to the other evaluation metrics. In Figure 2, the forecasts of seven different regression models for XAUUSD price are presented. Figure 2 demonstrates actual and predicted prices of XAUUSD employing different regression models. In each model, scatter plot is employed to observe the performance of different models for shuffled train and test data sets. It is obviously seen that the points estimated by the SR model for the test data almost completely coincide with the actual price.

V. DISCUSSION AND CONCLUSION

In this work, regression ensemble-based model is proposed to estimate the value of 1 ounce of gold in dollars (XAU/USD index). To demonstrate the effect of proposed regression ensemble-based model, seven different regression models namely, linear regression, polynomial regression, decision tree regression, random forest regression, support vector regression, voting regressor, stacking regressor are evaluated. For this purpose, the dataset is collected between July 2019 and July 2020 from financial websites, and enhanced with various indicators such as simple moving average (SMA) for 20, 50, and 100 days, opening, closing, highest and lowest dollar index (DXY) prices, 14-day relative strength index (RSI), the upper, middle and lower values of the Bollinger band (BB), and prepared for model construction. To the best of our knowledge, this is the very first attempt considering the consolidation of regression models for the estimation of XAU/USD index. Experiment results indicate that the utilization of stacking regression combination model demonstrates remarkable score with 2.2036 of MAPE for predicting the price of XAU/USD index. In future research, we plan to design a framework which blends both the results of time series analysis and regression models by adding various technical and statistical indicators into dataset.

FIG. 2. THE FORECASTS OF SEVEN DIFFERENT REGRESSION MODELS FOR XAUUSD PRICE

REFERENCES

[1]. Ismail, Z., Yahya, A., & Shabri, A. (2009). Forecasting gold prices using multiple linear regression method. American Journal of Applied Sciences, 6(8), 1509.

[2]. Manoj, J., & Suresh, K. K. (2019). Forecast Model for Price of Gold: Multiple Linear Regression with Principal Component Analysis. Thailand Statistician, 17(1), 125-131.

[3]. Jianwei, E., Ye, J., & Jin, H. (2019). A novel hybrid model on the prediction of time series and its application for the gold price analysis and forecasting. Physica A: Statistical Mechanics and Its Applications, 527, 121454.

[4]. Wei, Y., Liang, C., Li, Y., Zhang, X., & Wei, G. (2020). Can CBOE gold and silver implied volatility help to forecast gold futures volatility in China? Evidence based on HAR and Ridge regression models. Finance Research Letters, 35, 101287.

[5]. Pierdzioch, C., Risse, M., & Rohloff, S. (2015). A real-time quantile-regression approach to forecasting gold returns under asymmetric loss. Resources Policy, 45, 299-306.

[6]. Suranart, K., Kiattisin, S., & Leelasantitham, A. (2014, March). Analysis of comparisons for forecasting gold price using neural network, radial basis function network and support vector regression. In The 4th Joint International Conference on Information and Communication Technology, Electronic and Electrical Engineering (JICTEE) (pp. 1-5). IEEE.

[7]. Ongsritrakul, P., & Soonthornphisaj, N. (2003, July). Apply decision tree and support vector regression to predict the gold price. In Proceedings of the International Joint Conference on Neural Networks, 2003. (Vol. 4, pp. 2488-2492). IEEE.

[8]. Sadorsky, P. (2021). Predicting gold and silver price direction using tree-based classifiers. Journal of Risk and Financial Management, 14(5), 198.

[9]. M. A. Mithu, K. M. Rahman, R. A. Razu, M. Riajuliislam, S. I. Momo and A. Sattar. (2021, July). Gold price forecasting using regression techniques for settling economic and stock market inconsistency. In 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1-4, IEEE.

[10]. Yanto, M., Sanjaya, S., Yulasmi, Y., Guswandi, D., & Arlis, S. (2021). Implementation multiple linear regresion in neural network predict gold price. Indonesian Journal of Electrical Engineering and Computer Science, 22(3), 1635-1642.

[11]. Freedman, D. A. (2009). Statistical models: theory and practice. cambridge university press.

[12]. Rencher, A. C., & Christensen, W. F. (2012). Chapter 10, Multivariate regression–Section 10.1, Introduction. Methods of multivariate analysis, Wiley Series in Probability and Statistics, 709, 19.

[13]. Hilary, L. (1967). Seal. Studies in the history of probability and statistics. XV: The historical development of the Gauss linear model. Biometrika, 1-24.

[14]. Yan, X., & Su, X. (2009). Linear regression analysis: theory and computing. World Scientific.

[15]. Ostertagová, E. (2012). Modelling using polynomial regression. Procedia Engineering, 48, 500-506.

[16]. Quinlan, J. R. (1990). Decision trees and decision-making. IEEE Transactions on Systems, Man, and Cybernetics, 20(2), 339-346.

[17]. Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Angus, L. B., Yu, P. S., Zhou, Z., Steinberg, D. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14(1), 1-37.

[18]. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[19]. Korting, T. S. (2006). C4. 5 algorithm and multivariate decision trees. Image Processing Division, National Institute for Space Research–INPE Sao Jose dos Campos–SP, Brazil, 22.

[20]. Vapnik, V. N. (1999). An overview of statistical learning theory. IEEE transactions on neural networks, 10(5), 988-999.

[21]. Vapnik, V., Golowich, S., & Smola, A. (1996). Support vector method for function approximation, regression estimation and signal processing. Advances in neural information processing systems, 9.

[22]. An, K., & Meng, J. (2010, August). Voting-averaged combination method for regressor ensemble. In International Conference on Intelligent Computing (pp. 540-546). Springer, Berlin, Heidelberg.

[23]. Ruta, D., & Gabrys, B. (2005). Classifier selection for majority voting. Information fusion, 6(1), 63-81.

[24]. Džeroski, S., & Ženko, B. (2004). Is combining classifiers with stacking better than selecting the best one?. Machine learning, 54(3), 255-273.